

Article

Not peer-reviewed version

Exploring the Potential of Neural Machine Translation for Cross-Language Clinical NLP Resource Generation through Annotation Projection

[Jan Rodriguez Miret](#) , Eulalia Farre Maduell , Salvador Lima Lopez , [Laura Vigil Gimenez](#) , [Vicent Briva-Iglesias](#) , [Martin Krallinger](#) *

Posted Date: 8 August 2024

doi: 10.20944/preprints202408.0616.v1

Keywords: machine translation; annotation projection; clinical NLP; named entity recognition





Preprints.org is a free multidiscipline platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Article

Exploring the Potential of Neural Machine Translation for Cross-Language Clinical NLP Resource Generation through Annotation Projection

Jan Rodríguez-Miret ¹, Eulàlia Farré-Maduell ¹, Salvador Lima-López ¹ , Laura Vigil ¹, Vicent Briva-Iglesias ² and Martin Krallinger ^{1,*} 

¹ Barcelona Supercomputing Center (BSC)

² Dublin City University

* Correspondence: mkrallin@bsc.es; Tel.: (+34) 93 413 77 16

Abstract: Recent advancements in neural machine translation (NMT) offer promising potential for generating cross-language clinical natural language processing (NLP) resources. There is a pressing need to be able to foster the development of clinical NLP tools that extract key clinical entities in a comparable way for a multitude of medical application scenarios, hindered by lack of multilingual annotated data. This study explores the efficacy of using NMT and annotation projection techniques with expert in the loop validation to develop named entity recognition (NER) systems for an under resourced target language (Catalan) by leveraging Spanish clinical corpora annotated by domain experts. We employed a state-of-the-art NMT system to translate three clinical case corpora. The translated annotations were then projected onto the target language texts and subsequently validated and corrected by clinical domain experts. The efficacy of the resulting NER systems was evaluated against manually annotated test sets in the target language. Our findings indicate that this approach not only facilitates the generation of high-quality training data for the target language (Catalan) but also demonstrates the potential to extend this methodology to other languages, thereby enhancing multilingual clinical NLP resource development. The generated corpora and components are publicly accessible, providing potentially a valuable resource for further research and application in multilingual clinical settings: <https://zenodo.org/doi/10.5281/zenodo.13133124>.

Keywords: machine translation; annotation projection; clinical NLP; named entity recognition

1. Introduction

Recent advances in artificial intelligence (AI) technologies hold significant promise for the development of clinical natural language processing (NLP) applications [1]. These applications are crucial for the healthcare domain, where accurate and timely information extraction from clinical texts can significantly enhance patient care, research, and administrative processes [2–4]. Named entity recognition (NER) systems play a vital role by identifying and classifying key entities such as medications, symptoms, and diagnoses within clinical documents [5].

Clinical NER systems have demonstrated considerable impact. For instance, [6] showcased the use of NER to extract medical problems, treatments, and tests from clinical narratives, facilitating the creation of comprehensive patient summaries. Similarly, [7] analysed the use of NER systems to identify adverse drug events from clinical documents, improving pharmacovigilance and patient safety. These systems have helped optimise clinical workflows, enhancing decision support systems, and contributing to large-scale health data analysis.

However, the development of robust NER systems is contingent upon the availability of substantial annotated corpora. These corpora must be meticulously annotated by clinical experts to ensure accuracy, a process that is both time-consuming, costly, and entails extensive privacy risks [8]. Furthermore, annotated corpora are often language-specific, posing significant challenges in multilingual contexts where resources may be scarce for less widely spoken languages. Large clinical studies, the characterization of rare diseases or international medical projects do require that clinical information is extracted in a comparable manner, which is particularly challenging due to the lack of multilingual

clinical corpora annotated using similar data labelling criteria. Most of the current freely accessible resources are limited to data in English [9], and the use of clinical corpora in English as a base to generate resources in other languages and speed up NLP resource generation might show certain limitations due to the differences in syntax, grammar, and morphology, as well as lexical differences between languages [10].

It is also worth stressing that clinical content is typically written in the local language or even language variants spoken in each region or country [11]. In some geographical regions, clinical records might even correspond to textual narratives written in multiple languages, presenting therefore code-switching phenomena [12]. Such is the case of some Spanish regions, like Catalonia, where clinical health records are found both in Spanish and Catalan, resulting in multilingual content written in similar languages.

In this context, current advances in neural machine translation (NMT) might facilitate the creation of annotated corpora and the application of language technologies in multiple languages [13]. NMT systems can translate annotated corpora from one language to another while preserving the integrity and context of the original annotations [14,15]. This potential is enhanced within languages of the same family, like Romance languages, because of their similarity [16,17]. This study explores the use of a state-of-the-art NMT system (Softcatalà Translator) to translate three Spanish clinical case corpora into Catalan [18]. The original Spanish corpora, meticulously annotated by clinical experts, serve as the basis for this translation. To facilitate the transfer of annotations from Spanish to Catalan, we employ an annotation projection technique (explained in detail in Section 3). This method bypasses the need for extensive manual annotation of the Catalan corpora, thus saving significant time and resources.

This study details a strategy exploiting neural translation technologies, annotation projection and human in the loop expert validation to foster the development of comparable clinical NER systems and generation of annotated corpora together with a detailed evaluation and comparison of different NER systems. Under this application scenario, the use of NMT can effectively save time in manually annotating corpora. As a result, we have generated the first clinical NER components for three high impact entity types in Catalan. Through this study, we aim to provide insights into the feasibility and benefits of leveraging NMT and annotation projection for developing robust NER systems in multilingual clinical settings.

2. Neural Machine Translation for Annotation Projection

The advent of NMT in the past decade has revolutionized MT, utilizing deep learning techniques to enhance accuracy and contextual understanding [19,20]. This has translated into considerable effort over the past years to generate MT systems, not only for general domain purposes, but also for health-related documents [21–23]. Most medical MT efforts focused on translations between English and other languages [24], covering a broad variety of practical use case scenarios, such as the generation of medical dictionaries or glossaries, automatic clinical coding [25], processing of patient-physician interviews [26] or generating translation systems for low resource languages [22]. MT and word alignment strategies have been used between English and German to generate German NER models for medical semantic annotation of medication mentions [27,28], and for Spanish and French in the case of tumour morphology concepts [29].

Parallel and comparable biomedical corpora have also been explored to generate bilingual medical glossaries and terminological resources by exploiting different word alignment approaches and clinical vocabularies, for instance between content in English and French [30] or English and Spanish [31].

Despite recent advances in NMT technologies and the increasing use of large language models (LLMs) for translation-related tasks [32,33], it is surprising that only limited attempts have been made so far to exploit existing commonalities across medical texts of Romance languages [34]. In this paper, we propose a strategy to generate medical corpora more efficiently across languages by exploiting an NMT system for the development of clinical entity tagger resources.

The proposed approach relies on the use of NMT, followed by expert-validation of existing recently released and widely used Gold Standard clinical corpora available in Spanish to train and evaluate NER tools, in this case for Catalan. Figure 1 shows an overview of the proposed methodology. The sections below explain in detail the study conducted.

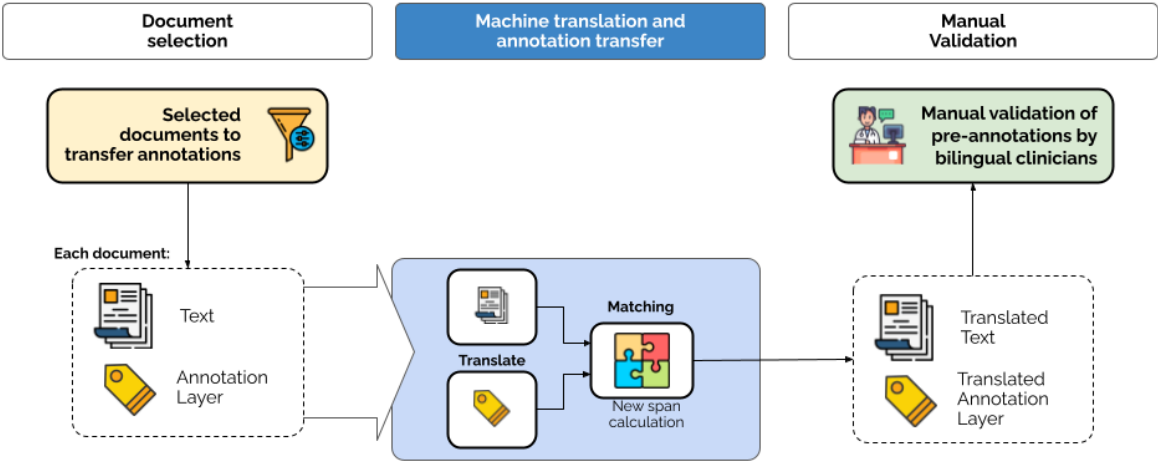


Figure 1. Overview of the clinical case report translation and entity annotation projection approach.

3. Materials and Methods

The number of corpora manually annotated by clinical experts available in Romance languages is scarce. This is particularly true for lower resource languages such as Catalan, Romanian or Galician. Recent community evaluation efforts and shared tasks have resulted in the release of a number of medical corpora in Spanish [35–38]. These have been widely exploited for the development and benchmarking of clinical NER systems [39]. This section describes the exploitation of three clinical corpora with NMT and annotation projection approaches to generate annotated resources for the Romance language Catalan.

3.1. Clinical Corpora: DisTEMIST, DrugTEMIST and MEDDOPROF

Three separate datasets were chosen for the annotation projection experiments, namely DrugTEMIST [40], DisTEMIST [41] and MEDDOPROF [42]. These are all Gold Standard corpora comprising clinical case reports written in Spanish that have been annotated manually by clinical domain experts. The annotations followed an iterative process in which different annotation guidelines were created and refined to ensure the quality and consistency of the final output. In addition, two of these corpora (DisTEMIST and MEDDOPROF) have been publicly validated as part of shared tasks.

The selection of multiple datasets with varying entity types allowed us to test the methodology’s effectiveness across a broad spectrum of clinical information, thereby providing a comprehensive assessment of its potential utility in multilingual clinical NLP applications. On the one hand, DisTEMIST and DrugTEMIST cover annotations of diseases and medications, respectively, as reported in 1,000 clinical case reports in Spanish per corpus that include medical, surgical and dentistry specialities. In the case of the DisTEMIST corpus, the annotation consistency measured through pairwise agreement between independent annotators reached 82.3% IAA (exact disease mention offsets). The DrugTEMIST corpus was annotated by the same team of clinical experts following detailed annotation guidelines. On the other hand, MEDDOPROF covers information on social determinants of health found in clinical texts (namely occupations and working statuses), an entity type that is not specific to the medical domain and that often appears in descriptive form. Table 1 provides some basic statistics for the three datasets.

Table 1. Characteristics of the three Gold Standard corpora used for the annotation projection experiment.

Dataset	Topic	Documents	Tokens	Annotations
DisTEMIST	Diseases	1,000	406,318	10,664
DrugTEMIST	Medications	1,000	406,318	2,782
MEDDOPROF	Occupations	1,844	1,291,186	4,770

3.2. Corpus Translation and Annotation Projection

In essence, the proposed annotation projection leverages labelled data or text-bound annotations from a source language (source corpus) into a different target language (target corpus). In the case of span-based annotations (i.e. for named entity recognition scenarios), this implies mapping the annotations of a given text in the source language into the translated text of the target language. For our experiments, we propose a *lexical* annotation projection, consisting of a more strict mapping strategy particularly focusing thus on human annotation correction efficiency as it decreases the need to correct matching boundary issues often encountered by more advanced sentence alignment-based matching alternatives. As opposed to previous approaches that mainly relied on sentence alignment techniques to detect correspondences between source and target concept mentions [30], the source text and the source annotations are translated separately into the target language and the annotations are then mapped onto the target text using term look-up methods. Some publicly released corpora that have tested this methodology include DisTEMIST [41], MEDDOPLACE [36], and SympTEMIST [37].

In the case of our approach, the translations from Spanish to Catalan were generated using the SoftCatala API¹. The choice of Softcatalà as the NMT system was influenced by several factors. First, Softcatalà Translator is an open-source system, which allows other researchers to freely access and replicate our methodology, contributing to the collective advancement in NMT for clinical NLP applications [43]. Second, ensuring the privacy of clinical data is paramount, and we could handle sensitive datasets without exposing them to potential security risks associated with other commercial NMT services. This compliance with ethical standards and legal regulations regarding patient data is crucial. Finally, Softcatalà Translator is renowned for its high-quality translations between Romance languages, displaying state-of-the-art results comparable with major NMT systems [18]. Maintaining the accuracy of translated annotations directly impacts the effectiveness of the annotation projection method and the performance of NER systems.

Consequently, the final output of the annotation projection process can be considered as a sort of Silver Standard corpus because the generated annotations in the target language text are not perfect and may contain missing mentions not matched or mapped to the target text, as well as incorrect annotations (either due to ambiguity as is often the case of abbreviations, or annotations with incorrect boundaries).

The quality of the annotation projection process is also dependent on the underlying NMT quality, the correctness of the source text, the medical expressions used, the potential ambiguity of the annotated entity mentions, or the similarity of the source and target languages. Even then, machine-translated corpora can still be quite valuable for multiple reasons, especially for under resourced languages. On the one hand, the automatically mapped annotations can be directly validated or corrected by native speakers with domain expertise to generate a new Gold Standard training or test set fast and efficiently. On the other hand, we wanted to test whether models trained on non-validated data (in a different language to that of the source corpora) can still provide sufficiently good results when evaluated on a test set that had been manually corrected by domain experts.

Table 2 shows an overview of the result of the annotation projection process of the three previously mentioned corpora from Spanish (seed language) to Catalan (target language). The statistics shown are

¹ <https://www.softcatala.org/traductor/>

calculated as part of the annotation projection process to give an estimate of the completeness of the projection, but they are not necessarily representative of the quality of the final output. Significantly, the majority of the Gold Standard annotations could be mapped onto the Catalan translations. This might be attributed to the closeness between both languages (Spanish and Catalan have 85% similarity) [44], the translation quality, as well as the existence of similar morphological, syntactic and even etymological patterns. This is especially relevant in the medical domain, where many clinical terms in both languages share the same Latin and Greek roots (DisTEMIST corpus). In the case of medication and drug names, these were almost always substance names with identical word roots for both languages (DrugTEMIST corpus). Since occupation mentions (MEDDOPROF corpus) sometimes correspond to more descriptive expressions, the percentage of non-projected annotations in this corpus is higher than for the other two corpora.

Table 2. Statistics of the projected corpora from Spanish to Catalan, divided in splits. GS stands for ‘Gold Standard’ (i.e. original Spanish corpus); Ann. stands for ‘annotations’.

Dataset	Documents	GS Ann.	Not Projected	% Not Projected
DisTEMIST TRAIN+DEV	750	8,065	98	1.21%
DisTEMIST TEST	250	2,599	55	2.11%
DrugTEMIST TRAIN+DEV	750	2,090	34	1.62%
DrugTEMIST TEST	250	692	18	2.60%
MEDDOPROF TRAIN+DEV	1,500	3,658	139	3.79%
MEDDOPROF TEST	344	1,085	54	4.97%

Crucially, and despite the seemingly good results, we should emphasize that without human validation it is not possible to truly assess the extent to which the resulting NMT and annotation projection are correct. It might be the case that the projected data includes false positives or incorrectly translated entities. For this reason, we incorporate a final step: human validation and correction of the annotation projection.

3.3. Corpus Validation and Correction Process

To be able to benchmark and evaluate the generated entity recognition results on systems trained on automatically projected annotations, it is critical to have a manually validated test set corrected by clinical domain experts.

For this experiment, two clinical experts, professional translator and one linguist were asked to validate the test set subsets of all three corpora. They had ample experience in text data annotation and are native Spanish and Catalan speakers. The validation step was performed using the annotation tool brat side-by-side document comparison mode [45]. In addition to the test set, to explore the difference between training systems using the annotation projection output as-is or using a corrected version, the experts were also asked to validate the train subsets of the corpora.

As shown by Figure 2, annotators were shown the Gold Standard annotated documents in Spanish on the left side and the projected document in Catalan on the right side. The main task for annotators was to validate the Catalan version, adding, removing or modifying annotations to make the projected version as close to the Gold Standard as possible. The annotation of a Gold Standard corpus from scratch requires annotators to carefully read the entire document and annotate mentions without any aids. In contrast, after NMT and annotation projection, a fast scan of the document usually suffices. The side-by-side modality further eases the annotators’ work by enabling direct comparisons between the source and target texts. This method significantly reduces the annotation effort, as annotators can focus on verifying and correcting the pre-existing annotations rather than creating them de novo.

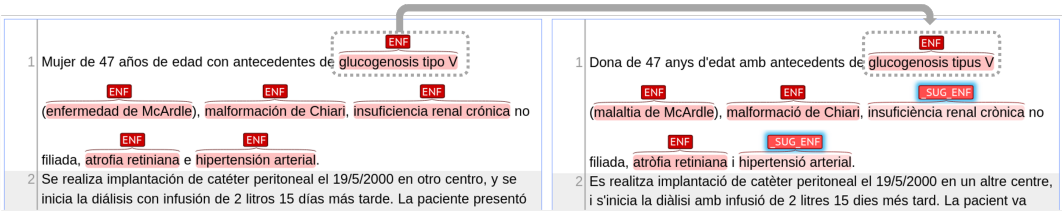


Figure 2. Example screenshot of BRAT interface side-by-side view used for the clinical expert validation and correction of entities. By double-clicking on each annotation’s label name, annotators were able to provide alternative translations using the Notes section.

A second task given to annotators was to verify that the annotated concepts were correctly translated and idiomatic. When the concept was incorrect, they were asked to suggest alternative translations. It is important to note that only the annotated concepts were corrected, since correcting the entire machine-translated text goes beyond the scope and point of the experiment. Annotators used the notes section in brat to propose improved translations for each individual annotation. During the data post-processing step, the translated texts were reconstructed. Here, incorrectly translated concepts were replaced by their proposed alternative, adjusting the character spans of all annotations as required.

Finally, in order to evaluate the resulting systems on non-machine-translated data, we gathered a set of clinical case reports in Catalan to be annotated by the three experts with the same set of labels (that is, diseases, medications and occupations). This data collection, named CataCCC, consists of case reports of various clinical specialties extracted from open journals. This method mimics the document selection used for the DisTEMIST and DrugTEMIST corpora. CataCCC includes a total of 200 documents with 72,003 tokens and 400,378 characters. It has 2,557 disease annotations, 531 medication annotations and 142 occupations. The corpus is openly available for download, along with the rest of the data discussed in this paper and the annotation projection correction guidelines, in Zenodo at <https://zenodo.org/doi/10.5281/zenodo.13133124>.

3.4. Clinical Named Entity Recognition Models

Transformer-based pre-trained language models have been increasingly used for NER tasks, also in the biomedical domain [46,47]. These models are trained on general tasks, like masked or causal language modelling, in a self-supervised manner. After that, the pre-trained models can be leveraged for different downstream tasks (e.g. text classification, NER, etc.) if properly fine-tuned using supervised data. To analyse the practical usefulness of the generated annotation projection corpora, we have trained several NER systems on the original Spanish corpora and the machine-translated Catalan corpora with the corrected mentions from the annotation projection. These systems were evaluated against human-annotated/validated test set annotations in all cases: (a) their corresponding Spanish test set, (b) the Catalan corrected test set, and (c) the CataCCC corpus, which is only used for evaluation. In this section, the developed NER models will be presented, along with the data used for their training and evaluation.

We formulate the task as a token classification problem, where each token is encoded using the IOB2 schema [48]. As a simple multi-class problem (all tokens must have exactly one class), in the case of nested mentions, only the longest one is kept for training. The model is, thus, unable to produce nested entities, but the percentage of these in the datasets is low, so the simplicity of this method outweighs the presumably negligible negative impact it may have. This approach, i.e., fine-tuning a pre-trained model on a token classification task, is also used by the top-scoring teams in the MEDDOPROF and DisTEMIST shared tasks [49,50].

As a base model, we used a RoBERTa-based model (124M parameters) pre-trained on both clinical and biomedical corpora, accounting for 91M tokens and 968M words, respectively [51]. We will refer to this model as RoBERTa bsc-bio-ehr-es. For each corpus, one-third of the original training set was selected as the validation or development set (250 documents for DisTEMIST and DrugTEMIST, and

500 documents for MEDDOPROF), which is used to evaluate the model at each epoch during training. The same splits are used for all experiment configurations and languages. For the loss computation, we used the cross-entropy loss implementation provided by Pytorch.

The label *ACTIVIDAD* (activity) was removed from the MEDDOPROF corpus after initial experimentation since the number of mentions is quite low (147 in the whole dataset) and they are semantically broad and complex. The rest of mentions and entities are kept as in the original corpora.

Notably, the dataset is mostly composed of *O* tags when encoded with the IOB2 schema, while some tags like *B-SITUACION_LABORAL* only account for 0.09% of tokens. To compensate for this imbalance, we tried different weight strategies for computing the loss:

- *none*: all tokens have the same importance.
- *freq*: each token has a weight inversely proportional to the frequency of its ground truth class (IOB tag) in the training split.

$$weight_{class_i} = \frac{ntokens}{ntokens_{class_i} \cdot nclasses} \quad (1)$$

where *ntokens* is the total number of tokens, *ntokens_{class_i}* the number of tokens for class *i*, and *nclasses* the number of classes.

- *freq_sqrt*: each token has a weight inversely proportional to the square root of the frequency of its ground truth class.

$$weight_{class_i} = \sqrt{\frac{ntokens}{ntokens_{class_i} \cdot nclasses}} \quad (2)$$

Note that both *freq* and the smoother *freq_sqrt* encourage a higher recall for entities, as false negatives (FN) are more penalized than false positives (FP).

3.4.1. Model Selection

For the MEDDOPROF corpus (Spanish), we also tested a general domain pre-trained model named BETO [52], but it was discarded after early experimentation in favor of the RoBERTa bsc-bio-ehr-es, which was consistently outperforming BETO. The models were trained in a HPC environment, where each node/experiment had 2 GPUs (AMD Radeon Pro VII, 16 GB each one), although their availability was not exclusively for the purpose of this study.

A hyperparameter optimization (HPO) exploration was conducted by means of the Weights & Biases Sweeps feature. We used the same sweep configuration for both Spanish and Catalan, running 15-30 random configurations for each corpus and language. The different possible values explored are listed below: **batch size**: {8, 16, 32}, **learning rate**: $[1^{-7}, 1^{-4}]$, **classifier dropout**: [0, 0.8], **weight decay**: [0, 0.03], **warmup ratio**: {0, 0.1}, **weight strategy**: {*none*, *freq*, *freq_sqrt*}, **number of epochs**: {15, 20}.

The best model is selected as the one with the highest strict mention-wise F1-score at some epoch *k* in the validation set. There are two exceptions to this: in both Spanish DisTEMIST and MEDDOPROF, the model with the highest F1-score did not apply the dropout regularization technique (i.e. classifier dropout = 0). The difference in performance with respect to the second-highest F1-score was less than 1 percentage point in both cases, but these had the dropout set to a value of 0.3 and 0.8, respectively. We hypothesized that these latter models would result in a better generalization and thus a higher performance in the test set. After all experiments, we evaluated the former two discarded models in the test set and confirmed that their performance was 1.2 percentage points below the chosen regularized ones.

3.4.2. Model Availability

To ensure transparency and reproducibility, and encourage further research on the use of NMT and annotation projection in languages other than English, all models appearing in this article are publicly available at Hugging Face: <https://huggingface.co/collections/BSC-NLP4BIA/clinical-nmt->

[ner-66a7badd9189298acbbe504f](#). More details on each model, license, and how to use them can be found on each Model Card.

As initial baseline, we trained models on the original annotation projection data without human correction. These are noted with the suffix “-v0” in the above Hugging Face collection. Associated details are provided as additional materials in Zenodo.

4. Results

4.1. Cross-Language Model Evaluation

To evaluate the transferability of the Spanish NER corpora to Catalan, we trained a transformer-based model on each of the original Spanish training set corpus and on the corrected training set version of the Catalan automatically transferred corpus.

All results can be seen in Table 3. As expected, the best-performing models in both the Catalan test set and CataCCC are the ones trained on the Catalan corrected version of the corpora, while the models trained on the original Spanish corpora achieved the best metrics in their corresponding test sets. Interestingly, the Spanish model for drugs obtained a higher precision than the Catalan corrected version when evaluated on the test set of the latter (0.886 vs 0.885), though with a much lower recall (0.809 vs 0.874).

Note that the models trained and evaluated on the original Spanish versions set a higher bound or reference in performance regarding the Catalan version. That is, it was not expected to surpass these performances for the settings trained and evaluated on the Catalan version, as the RoBERTa base model was trained on Spanish texts.

Table 3. Results for the test sets of DrugTEMIST, DisTEMIST, and MEDDOPROF (“Spa” and “Cat”), and CataCCC. Models were trained on the training set of the original Spanish version (“Spa”) and the corrected Catalan version (“Cat corr.”). Best metrics for each corpus are marked in bold. Note that each square represents the results of the same model evaluated on the different corpora. Only the best/selected model from all hyperparameter optimization experiments is shown.

Trained on/Eval. on	Drugs			Diseases			Occupations		
	P	R	F1	P	R	F1	P	R	F1
Spa/Spa	0.917	0.909	0.913	0.754	0.759	0.757	0.785	0.776	0.780
Spa/Cat corr.	0.886	0.809	0.846	0.610	0.608	0.609	0.642	0.337	0.442
Spa/CataCCC	0.908	0.857	0.882	0.742	0.702	0.721	0.750	0.444	0.558
Cat corr./Spa	0.884	0.880	0.882	0.672	0.644	0.658	0.723	0.726	0.725
Cat corr./Cat corr.	0.885	0.874	0.879	0.701	0.718	0.709	0.743	0.717	0.729
Cat corr./CataCCC	0.921	0.904	0.913	0.775	0.818	0.796	0.838	0.794	0.815

It is important to emphasize the fact that F1 scores increase largely when not considering the exact boundary of a mention but just some overlap between the prediction and the ground truth. This is especially true for DisTEMIST, which has the longest mentions (see Table 4).

Table 4. Strict and relaxed F1-score using the MEDDOPLACE Scoring Script [36] for . The relaxed metric will count a prediction as a true positive if it has at least some overlap with a ground truth mention.

Trained on/ Eval. on	Drugs		Diseases		Occupations	
	Strict	Relaxed	Strict	Relaxed	Strict	Relaxed
Spa/Spa	0.913	0.954	0.757	0.892	0.782	0.872
Cat corr./Cat corr.	0.879	0.937	0.709	0.866	0.729	0.840
Cat corr./CataCCC	0.913	0.957	0.796	0.889	0.815	0.875

4.2. MT Error Analysis

Even if our primary focus is on using NMT for data generation in clinical NLP through an annotation projection technique, understanding how MT errors can impact the quality and accuracy

of the projected annotations is crucial. The three expert annotators assessed the MT quality by using the error categories in the Multidimensional Quality Metrics (MQM) framework for human evaluation of translation quality, namely Terminology, Accuracy, Linguistic Conventions and Style [53]. Following this strict methodology [54,55], we identified areas for MT improvement, which impacted the annotation projection. The MT quality analysis was conducted separately for each corpus, since the impact the MT results had on the annotation projection was different for each of the 3 entities (diseases, medications and occupations).

The DisTEMIST corpus contains a high number of annotated disease entities per document. Diseases have specialized, in-domain terminology and, as expected, most MT errors in this corpus fall into the Terminology and Accuracy categories because the MT system failed to adequately translate some of these entities. One common Terminology error identified was the MT system producing literal translations, such as the case of “enfermedad de 3 vasos” (disease affecting 3 vessels), which was machine translated as “malaltia de 3 gots” (disease of 3 glasses) instead of the adequate “malaltia de 3 vasos”, mixing up key concepts. In terms of Accuracy, one common error was not translating the entity at all, like leaving in Spanish “bifidez piélica” (bifid pelvis) instead of translating the entity into Catalan as “bífidesa pièlica”. Regarding Linguistic Conventions, there were some typos, as in “metástasis” instead of “metàstasi”. Surprisingly, the quality of the MT generally did not hinder the annotation transfer in this corpus. Despite the MT errors, the robustness of the annotation projection technique ensured that all the above disease annotations were accurately transferred.

The DrugTEMIST corpus, which contains annotations of medicines, presented unique challenges. Proper nouns and specific drug names were particularly prone to errors, often falling into the Terminology and Accuracy categories. For instance, the MT system frequently mistranslated drug names by either altering their endings or failing to detect them as proper nouns. A common Terminology error was translating nouns as verbs, such as in the case of “sales de fosfato” (phosphate salts) as “surts de fosfat” (you leave phosphate). Accuracy errors included cases where the MT system did not translate the term at all, leaving “ácido fólnico” untranslated instead of converting it to “àcid folínic”. Additionally, there were instances of inappropriate translation of abbreviations and symbols, such as “5-ASA” being incorrectly translated to “5-NANSA” (in Spanish, “asa” can refer to mesalazine in the medical domain, or to a bag handle, which is the option that the MT system chose). These errors highlight the need for meticulous verification of medical terminology in MT systems to ensure accurate annotation projection.

Finally, the MEDDOPROF corpus, which deals with occupations, demonstrated different types of MT errors. Given the descriptive nature of the annotations of this latter corpus, the primary issues were related to Terminology and Linguistic Conventions. Abbreviations posed significant challenges; for example, “estudiante de 3º ESO” (student of 3rd year of secondary school) was often incorrectly translated to “estudiant de 3º ESO” instead of “estudiant de 3r ESO”. The error in the translation of the abbreviation caused that every annotation containing an abbreviation was not properly transferred to the Catalan corpus. Terminology errors included the mistranslation of terms such as “médico” (doctor) to “mèdic” (medical) rather than the correct “metge” (doctor). Another similar case was the mistranslation of “empleado en Carpintería metálica” (worker in metallic carpentry) to “emprat en Carpinteria metàl·lica” (used in metallic carpentry). In these cases, the annotation was not projected adequately. These errors underline the importance of high-quality MT systems for accurate annotation projection, especially in clinical contexts where precision is crucial. Further implications on the MT are addressed in Section 5.

4.3. Annotation Projection and NER System Error Analysis

In this section, we analyse the errors encountered during the annotation projection and the subsequent performance of the NER systems trained on the projected annotations. This analysis aims to identify the common sources of errors and their potential impacts on the system’s overall performance.

In the qualitative analysis of the annotations, we could observe some false positives and false negatives. The most prevalent type of false positive error involved substances that could be both medicinal products and physiological substances, such as *potassium* and *albumin*. In the original Spanish version, these were only annotated when referring to medicinal products. However, in the Catalan version, the annotation projection also considered these concepts when describing blood test results, leading to over-annotation. False negatives were most commonly observed with medicinal products that included specific concentrations, such as *sodium chloride 0.9% solution*. These cases often required manual annotation in the target language due to incomplete projection.

Using the CataCCC corpus for evaluation, we classified model predictions into correct, incorrect, partial, missing, and spurious categories to gain a more profound understanding of performance. This methodology was employed by following the MUC-5 evaluation [56]. We observed that incorrect boundaries were a frequent issue, often due to variations in the level of detail in the annotations.

For drug annotations (DrugTEMIST), the model faced challenges with special characters like the Catalan “l·l”, apostrophes, and other symbols such as percentages. These issues are likely related to inadequate tokenization for these characters. For example, *al·lopurinol* and *metil·lina* were often incorrectly tokenized, affecting the model’s ability to correctly identify these terms.

In the DisTEMIST corpus, many predictions included extended descriptions or adjectives that were not present in the Gold Standard corpus, reflecting the difficulty in standardizing the level of detail for disease mentions. Additionally, many spurious (totally false positive) and missing (totally false negative) mentions can also be considered symptoms, which illustrates the blurred line between the two semantic classes.

For occupations (MEDDOPROF), incorrect boundary predictions were common, often due to varying levels of descriptive detail. Some mentions were particularly difficult for the model to identify as occupations or employment statuses, given their use in common language rather than specific occupational terms. For example, “diagnostic radiology technician” and “on disability benefits” required precise context understanding, which was often lacking.

5. Discussion

Overall, the results obtained indicate that the use of NMT, together with an annotation projection technique, might be used to generate medical NLP training data. We achieved competitive results for our target language when compared with the source language. Moreover, we have tested this strategy for automatic clinical semantic annotation focusing on three distinct relevant entity types, namely medications, diseases, and occupations. Each entity type has specific linguistic characteristics in terms of MT results when considering the source and target languages of our experiment. Medications often correspond to short substance or trade name mentions with certain morphological characteristics, suffixes and prefixes that follow an established pattern between the two Romance languages of our study. In the case of disease mentions, these often consisted of complex, long multi-word expressions built up by words with Latin or Greek roots. Occupations, on the other hand, are not exclusive to medical content. Nevertheless, when encountered in clinical narratives, they usually appear both in the form of proper occupation names or as more descriptive expressions.

When using transformer-based clinical NER systems with biomedical RoBERTa, we observed that the performance only degraded slightly when the system was trained on the annotation projection dataset and applied to another similar language, i.e., Catalan. The best results were obtained for medication mentions.

Our results indicate that the reuse of existing clinical corpora for similar language adaptation scenarios could be a promising strategy to accelerate the annotation process and generate resources of importance under data scarcity, as is the case of clinical corpora. In our case, due to the existence of clinical data written in both Spanish and Catalan, accounting for a population of over 10 million people, unlocking clinically relevant information requires systems that work well for content in either language. For this reason, machine translation of texts from Spanish to Catalan for the development

of NLP resources and systems in Catalan might considerably enhance patient comprehension and communication by providing medical information in their native language within hospitals. It could also improve operational efficiency and ensure consistent medical terminology, although it requires careful review for accuracy and privacy compliance.

The three steps of the corpus construction experiment, namely, annotation projection, projection validation and correction of translated mentions, have underscored important factors to be considered when leveraging resources between different languages. Firstly, improvements in the quality of the MT system would result in a better projection of annotations. The corpora written mainly in Castilian Spanish (DisTEMIST and DrugTEMIST) seemed to have somewhat better translation results when compared to the panhispanic MEDDOPROF corpus, which covered clinical cases not only from Spain but also a diversity of Latin American countries. Efforts are being made to record the homogeneity and heterogeneity of the medical language shared by all Spanish speakers covering the entire human geography and language variants but are currently limited to medical terminologies rather than MT aspects [57], but a fine-tuning of MT systems in the medical domain may facilitate these tasks [58].

Aspects related to language variants and adaptation of annotation projection and automatic semantic annotation systems to similar languages would require a more in-depth analysis to determine the impact and issues to be addressed. An accurate translation is needed not only for the entity mentions *per se*, but also for the grammatical elements dependent on it, since Romance languages use declensions universally. Other issues are associated with ambiguous words and abbreviations in the original language, which are sometimes translated incorrectly in the target language. For example, in Spanish *ampolla* means ampule and blister (*ampul·la* and *butllofa* in Catalan, respectively).

The system used for annotation projection also showed some inconsistencies. For instance, the Spanish medication mention *amoxicilina-àcid clavulànic* was found sometimes correctly translated into *amoxicil·lina-àcid clavulànic*, and in other instances projected directly from Spanish without changes. Additionally, the resulting projection did not account for the actual context and annotation semantics of a particular mention. In the original Spanish text, *magnesium* was correctly labelled only when used as a medication, while in the target language it was sometimes incorrectly transferred where it corresponded to the physiological ion (semantic discrepancies). The projection of annotations was more challenging for entities that consist of noun phrases. One example is the case of occupations and work statuses, *diagnostic radiology technician* and *on disability benefits*, respectively. Regarding diseases, the length of the mention can vary greatly, particularly with the addition of anatomical parts affected by the condition. Examples of shorter and longer mentions are *brucellosis* and *right sphenoid wing en plaque meningioma*.

6. Conclusions

We have shown that annotation projection via NMT, human expert correction and transformer-based NER can yield promising results to generate resources for clinical NLP for similar languages scenarios under the assumption that high-quality MT systems are in place. We have also made available one of the first clinical NER Gold Standard corpora and systems of diseases, drugs and professions in Catalan. These information types, especially drugs and diseases, are very high-impact information types in a clinical setting, and the models trained for them achieve very competitive results when compared to their Spanish counterpart.

The proposed approach could potentially be explored also for other currently accessible clinical corpora in Spanish (e.g. CANTEMIST [59], PharmaCoNER [60], MEDDOPLACE [36], MedProcNER [38] or SympTEMIST [37]) and languages similar to Spanish like Portuguese (both European and Brazilian) or Galician. As MT resources from Spanish into multiple Romance languages are currently available, including also French, Italian, Romanian, Corsican or Haitian Creole, this could allow for systematically testing the proposed strategy with languages showing different degrees of similarity. So far we have generated as additional annotation projection corpora versions of the described datasets for English, French, Italian, Portuguese, Romanian, Czech, Dutch and Swedish.

Furthermore, to address data augmentation, the use of different MT systems or back-translation as well as the use of generative models and synthetic clinical texts could be tested to increase the number of training instances with particular focus on entity-aware NMT results.

Finally, it would be interesting to explore how the proposed setting would work for other application domains, genres, or content types such as social media or user/patient generated content.

Author Contributions: Conceptualization, M.K.; methodology, J.R.; software, X.X.; validation, E.F.; L.V.; V.B.; formal analysis, X.X.; investigation, J.R.; E.F.; S.L.; L.V.; B.V.; M.K resources, X.X.; data curation, S.L.; E.F.; L.V.; V.B. writing—original draft preparation, S.L.; M.K. writing—review and editing, S.L.; E.F.; M.K. visualization, J.R.; supervision, M.K.; project administration, M.K.; funding acquisition, Y.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Ministerio de Ciencia e Innovación (MICINN) under project AI4ProfHealth (PID2020-119266RA-I00 MICIU/AEI/10.13039/501100011033) and BARITONE (TED2021-129974B-C22). This work is also supported by the European Union's Horizon Europe Co-ordination & Support Action under Grant Agreement No 101080430 (AI4HF) as well as Grant Agreement No 101057849 (DataTool4Heartproject).

Data Availability Statement: The original data presented in the study are openly available in Zenodo at <https://zenodo.org/doi/10.5281/zenodo.13133124>.

Acknowledgments: We acknowledge Miguel Rodriguez Ortega, Luis Gasco Sanchez and Darryl Johan Estrada for collaborations and previous work related to neural machine translation and biomedical WMT shared task activities relevant for some of the initial experiments that motivated this work.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Al Kuwaiti, A.; Nazer, K.; Al-Reedy, A.; Al-Shehri, S.; Al-Muhanna, A.; Subbarayalu, A.V.; Al Muhanna, D.; Al-Muhanna, F.A. A Review of the Role of Artificial Intelligence in Healthcare. *Journal of Personalized Medicine* **2023**, *13*, 951. <https://doi.org/10.3390/jpm13060951>.
2. Houssein, E.H.; Mohamed, R.E.; Ali, A.A. Machine Learning Techniques for Biomedical Natural Language Processing: A Comprehensive Review. *IEEE Access* **2021**, *9*, 140628–140653. <https://doi.org/10.1109/ACCESS.2021.3119621>.
3. Kundeti, S.R.; Vijayananda, J.; Mujjiga, S.; Kalyan, M. Clinical Named Entity Recognition: Challenges and Opportunities. In Proceedings of the 2016 IEEE International Conference on Big Data (Big Data), 2016, pp. 1937–1945. <https://doi.org/10.1109/BigData.2016.7840814>.
4. Pagad, N.S.; Pradeep, N. Clinical Named Entity Recognition Methods: An Overview. In Proceedings of the International Conference on Innovative Computing and Communications; Khanna, A.; Gupta, D.; Bhattacharyya, S.; Hassanien, A.E.; Anand, S.; Jaiswal, A., Eds., Singapore, 2022; pp. 151–165. https://doi.org/10.1007/978-981-16-2597-8_13.
5. Wu, Y.; Jiang, M.; Xu, J.; Zhi, D.; Xu, H. Clinical Named Entity Recognition Using Deep Learning Models. *AMIA Annual Symposium Proceedings* **2018**, *2017*, 1812–1819.
6. Uzuner, Ö.; South, B.R.; Shen, S.; DuVall, S.L. 2010 I2b2/VA Challenge on Concepts, Assertions, and Relations in Clinical Text. *Journal of the American Medical Informatics Association : JAMIA* **2011**, *18*, 552–556. <https://doi.org/10.1136/amiajnl-2011-000203>.
7. Luo, Y.; Thompson, W.K.; Herr, T.M.; Zeng, Z.; Berendsen, M.A.; Jonnalagadda, S.R.; Carson, M.B.; Starren, J. Natural Language Processing for EHR-Based Pharmacovigilance: A Structured Review. *Drug Safety* **2017**, *40*, 1075–1089. <https://doi.org/10.1007/s40264-017-0558-6>.
8. Hovy, D.; Prabhumoye, S. Five Sources of Bias in Natural Language Processing. *Language and Linguistics Compass* **2021**, *15*, e12432. <https://doi.org/10.1111/lnc3.12432>.
9. Névél, A.; Dalianis, H.; Velupillai, S.; Savova, G.; Zweigenbaum, P. Clinical natural language processing in languages other than English: opportunities and challenges. *Journal of biomedical semantics* **2018**, *9*, 1–13.
10. Schneider, E.T.R.; de Souza, J.V.A.; Knafo, J.; e Oliveira, L.E.S.; Copara, J.; Gumiel, Y.B.; de Oliveira, L.F.A.; Paraiso, E.C.; Teodoro, D.; Barra, C.M.C.M. BioBERTpt - A Portuguese Neural Language Model for Clinical Named Entity Recognition. In Proceedings of the Proceedings of the 3rd Clinical Natural Language Processing Workshop; Rumshisky, A.; Roberts, K.; Bethard, S.; Naumann, T., Eds., Online, 2020; pp. 65–72. <https://doi.org/10.18653/v1/2020.clinicalnlp-1.7>.

11. García-Izquierdo, I.; Montalt, V. Cultural Competence and the Role of the Patient's Mother Tongue: An Exploratory Study of Health Professionals' Perceptions. *Societies* **2022**, *12*, 1735–1780.
12. Montalt, V. Ethical Considerations in the Translation of Health Genres in Crisis Communication. In *Translating Crises*; Bloomsbury Publishing, 2022; pp. 17–36.
13. Schäfer, H.; Idrissi-Yaghir, A.; Horn, P.; Friedrich, C. Cross-Language Transfer of High-Quality Annotations: Combining Neural Machine Translation with Cross-Linguistic Span Alignment to Apply NER to Clinical Texts in a Low-Resource Language. In Proceedings of the Proceedings of the 4th Clinical Natural Language Processing Workshop; Naumann, T.; Bethard, S.; Roberts, K.; Rumshisky, A., Eds., Seattle, WA, 2022; pp. 53–62. <https://doi.org/10.18653/v1/2022.clinicalnlp-1.6>.
14. Xie, S.; Xia, Y.; Wu, L.; Huang, Y.; Fan, Y.; Qin, T. End-to-End Entity-Aware Neural Machine Translation. *Machine Learning* **2022**, *111*, 1181–1203. <https://doi.org/10.1007/s10994-021-06073-9>.
15. Jain, A.; Paranjape, B.; Lipton, Z.C. Entity Projection via Machine Translation for Cross-Lingual NER, 2019, [arXiv:cs, stat/1909.05356]. <https://doi.org/10.48550/arXiv.1909.05356>.
16. Mikolov, T.; Le, Q.V.; Sutskever, I. Exploiting Similarities among Languages for Machine Translation, 2013, [arXiv:cs/1309.4168]. <https://doi.org/10.48550/arXiv.1309.4168>.
17. Altintas, K.; Cicekli, I. A Machine Translation System Between a Pair of Closely Related Languages. In *International Symposium on Computer and Information Sciences*; CRC Press, 2002.
18. Briva-Iglesias, V. English-Catalan Neural Machine Translation: State-of-the-Art Technology, Quality, and Productivity. *Tradumàtica: tecnologies de la traducció* **2022**, pp. 149–176. <https://doi.org/10.5565/rev/tradumatica.303>.
19. Castilho, S.; Moorkens, J.; Gaspari, F.; Calixto, I.; Tinsley, J.; Way, A. Is Neural Machine Translation the New State of the Art? *The Prague Bulletin of Mathematical Linguistics* **2017**.
20. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. *Advances in neural information processing systems* **2017**, *30*.
21. Costa-Jussà, M.R.; Farrús, M.; Serrano Pons, J. Machine translation in medicine: A quality analysis of statistical machine translation in the medical domain **2012**.
22. Soto, X.; Perez-de Viñaspre, O.; Oronoz, M.; Labaka, G. Development of a machine translation system for promoting the use of a low resource language in the clinical domain: The case of Basque. In *Natural Language Processing In Healthcare*; CRC Press, 2022; pp. 139–158.
23. Soares, F.; Krallinger, M. BSC participation in the WMT translation of biomedical abstracts. In Proceedings of the Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2), 2019, pp. 175–178.
24. Neves, M.; Yepes, A.J.; Siu, A.; Roller, R.; Thomas, P.; Navarro, M.V.; Yeganova, L.; Wiemann, D.; Di Nunzio, G.M.; Vezzani, F.; et al. Findings of the WMT 2022 biomedical translation shared task: Monolingual clinical case reports. In Proceedings of the WMT22-Seventh Conference on Machine Translation, 2022, pp. 694–723.
25. Almagro, M.; Martínez, R.; Montalvo, S.; Fresno, V. A cross-lingual approach to automatic ICD-10 coding of death certificates by exploring machine translation. *Journal of biomedical informatics* **2019**, *94*, 103207.
26. Pilegaard, M. Translation of medical research articles. *Benjamins Translation Library* **1997**, *26*, 159–184.
27. Frei, J.; Frei-Stuber, L.; Kramer, F. GERNERMED++: Semantic annotation in German medical NLP through transfer-learning, translation and word alignment. *Journal of Biomedical Informatics* **2023**, *147*, 104513.
28. Schäfer, H.; Idrissi-Yaghir, A.; Horn, P.; Friedrich, C. Cross-Language Transfer of High-Quality Annotations: Combining Neural Machine Translation with Cross-Linguistic Span Alignment to Apply NER to Clinical Texts in a Low-Resource Language. In Proceedings of the Proceedings of the 4th Clinical Natural Language Processing Workshop; Naumann, T.; Bethard, S.; Roberts, K.; Rumshisky, A., Eds., Seattle, WA, 2022; pp. 53–62. <https://doi.org/10.18653/v1/2022.clinicalnlp-1.6>.
29. Zaghir, J.; Bjelogrić, M.; Goldman, J.P.; Aananou, S.; Gaudet-Blavignac, C.; Lovis, C. FRASIMED: a Clinical French Annotated Resource Produced through Crosslingual BERT-Based Annotation Projection. *arXiv preprint arXiv:2309.10770* **2023**.
30. Deléger, L.; Merkel, M.; Zweigenbaum, P. Translating medical terminologies through word alignment in parallel text corpora. *Journal of Biomedical Informatics* **2009**, *42*, 692–701.
31. Villegas, M.; Intxaurreondo, A.; Gonzalez-Agirre, A.; Marimon, M.; Krallinger, M. The MeSpEN resource for English-Spanish medical machine translation and terminologies: census of parallel corpora, glossaries and term translations. *LREC MultilingualBIO: multilingual biomedical text processing* **2018**.

32. Yang, R.; Tan, T.F.; Lu, W.; Thirunavukarasu, A.J.; Ting, D.S.W.; Liu, N. Large language models in health care: Development, applications, and challenges. *Health Care Science* **2023**, *2*, 255–263.
33. Briva-Iglesias, V.; Camargo, J.L.C.; Dogru, G. Large Language Models "Ad Referendum": How Good Are They at Machine Translation in the Legal Domain?, 2024, [arXiv:cs.CL/2402.07681].
34. ten Hacken, P. The Language of Medicine in the Romance Languages. In *Oxford Research Encyclopedia of Linguistics*; 2023.
35. Miranda-Escalada, A.; Farré-Maduell, E.; Lima-López, S.; Estrada, D.; Gascó, L.; Krallinger, M. Mention detection, normalization & classification of species, pathogens, humans and food in clinical documents: Overview of LivingNER shared task and resources. *Procesamiento del Lenguaje Natural* **2022**.
36. Lima-López, S.; Farré-Maduell, E.; Brivá-Escalada, V.; Gascó, L.; Krallinger, M. MEDDOPLACE Shared Task overview: recognition, normalization and classification of locations and patient movement in clinical texts. *Procesamiento del Lenguaje Natural* **2023**, 71.
37. Lima-López, S.; Farré-Maduell, E.; Gasco-Sánchez, L.; Rodríguez-Miret, J.; Krallinger, M. Overview of SympTEMIST at BioCreative VIII: corpus, guidelines and evaluation of systems for the detection and normalization of symptoms, signs and findings from text. In Proceedings of the Proceedings of the BioCreative VIII Challenge and Workshop: Curation and Evaluation in the era of Generative Models, 2023.
38. Lima-López, S.; Farré-Maduell, E.; Gascó, L.; Nentidis, A.; Krithara, A.; Katsimpras, G.; Paliouras, G.; Krallinger, M. Overview of MedProcNER task on medical procedure detection and entity linking at BioASQ 2023. In Proceedings of the Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, 2023.
39. Shaitarova, A.; Zaghir, J.; Lavelli, A.; Krauthammer, M.; Rinaldi, F. Exploring the Latest Highlights in Medical Natural Language Processing across Multiple Languages: A Survey. *Yearbook of medical informatics* **2023**, *32*, 230–243.
40. Lima-López, S.; Farré-Maduell, E.; Rodríguez-Miret, J.; Rodríguez-Ortega, M.; Lilli, L.; Lenkowicz, J.; Ceroni, G.; Kossoff, J.; Shah, A.; Nentidis, A.; et al. Overview of MultiCardioNER task at BioASQ 2024 on Medical Speciality and Language Adaptation of Clinical NER Systems for Spanish, English and Italian. In Proceedings of the CLEF Working Notes; Faggioli, G.; Ferro, N.; Galuščáková, P.; García Seco de Herrera, A., Eds., 2024.
41. Miranda-Escalada, A.; Gascó, L.; Lima-López, S.; Farré-Maduell, E.; Estrada, D.; Nentidis, A.; Krithara, A.; Katsimpras, G.; Paliouras, G.; Krallinger, M. Overview of DisTEMIST at BioASQ: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources **2022**.
42. Lima-López, S.; Farré-Maduell, E.; Miranda-Escalada, A.; Brivá-Iglesias, V.; Krallinger, M. NLP applied to occupational health: MEDDOPROF shared task at IberLEF 2021 on automatic recognition, classification and normalization of professions and occupations from medical texts. *Procesamiento del Lenguaje Natural* **2021**, *67*, 243–256.
43. Team, N.; Costa-jussà, M.R.; Cross, J.; Çelebi, O.; Elbayad, M.; Heafield, K.; Heffernan, K.; Kalbassi, E.; Lam, J.; Licht, D.; et al. No Language Left Behind: Scaling Human-Centered Machine Translation, 2022, [arXiv:cs/2207.04672]. <https://doi.org/10.48550/arXiv.2207.04672>.
44. Ethnologue. The Catalan Language. <https://www.ethnologue.com/language/cat/>, 2017.
45. Stenetorp, P.; Pyysalo, S.; Topić, G.; Ohta, T.; Ananiadou, S.; Tsujii, J. BRAT: a web-based tool for NLP-assisted text annotation. In Proceedings of the Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, 2012, pp. 102–107.
46. Tian, S.; Erdengasileng, A.; Yang, X.; Guo, Y.; Wu, Y.; Zhang, J.; Bian, J.; He, Z. Transformer-Based Named Entity Recognition for Parsing Clinical Trial Eligibility Criteria. In Proceedings of the Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics, New York, NY, USA, 2021; BCB '21, pp. 1–6. <https://doi.org/10.1145/3459930.3469560>.
47. Yang, F.; Wang, X.; Ma, H.; Li, J. Transformers-Sklearn: A Toolkit for Medical Language Understanding with Transformer-Based Models. *BMC Medical Informatics and Decision Making* **2021**, *21*, 90. <https://doi.org/10.1186/s12911-021-01459-0>.
48. Ratnaparkhi, A.; Marcus, M.P. Maximum entropy models for natural language ambiguity resolution. PhD thesis, USA, 1998. AAI9840230.
49. Lange, L.; Adel, H.; Strötgen, J. Boosting Transformers for Job Expression Extraction and Classification in a Low-Resource Setting, 2021. arXiv:2109.08597 [cs] version: 1, <https://doi.org/10.48550/arXiv.2109.08597>.

50. Moscato, V.; Postiglione, M.; Sperli, G. Biomedical Spanish Language Models for entity recognition and linking at BioASQ DisTEMIST. 09 2022.
51. Carrino, C.P.; Llop, J.; Pàmies, M.; Gutiérrez-Fandiño, A.; Armengol-Estapé, J.; Silveira-Ocampo, J.; Valencia, A.; Gonzalez-Agirre, A.; Villegas, M. Pretrained Biomedical Language Models for Clinical NLP in Spanish. In Proceedings of the Proceedings of the 21st Workshop on Biomedical Language Processing; Demner-Fushman, D.; Cohen, K.B.; Ananiadou, S.; Tsujii, J., Eds., Dublin, Ireland, 2022; pp. 193–199. <https://doi.org/10.18653/v1/2022.bionlp-1.19>.
52. Cañete, J.; Chaperon, G.; Fuentes, R.; Ho, J.H.; Kang, H.; Pérez, J. Spanish Pre-Trained BERT Model and Evaluation Data. In Proceedings of the PML4DC at ICLR 2020, 2020.
53. Lommel, A. Metrics for Translation Quality Assessment: A Case for Standardising Error Typologies. In *Translation Quality Assessment: From Principles to Practice*; Moorkens, J.; Castilho, S.; Gaspari, F.; Doherty, S., Eds.; Springer International Publishing: Cham, 2018; pp. 109–127. https://doi.org/10.1007/978-3-319-91241-7_6.
54. Freitag, M.; Foster, G.; Grangier, D.; Ratnakar, V.; Tan, Q.; Macherey, W. Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. *Transactions of the Association for Computational Linguistics* **2021**, *9*, 1460–1474.
55. Moorkens, J.; Castilho, S.; Gaspari, F.; Doherty, S., Eds. *Translation Quality Assessment: From Principles to Practice*; Vol. 1, *Machine Translation: Technologies and Applications*, Springer International Publishing: Cham, 2018. <https://doi.org/10.1007/978-3-319-91241-7>.
56. Chinchor, N.; Sundheim, B. MUC-5 Evaluation Metrics. In Proceedings of the Fifth Message Understanding Conference (MUC-5): Proceedings of a Conference Held in Baltimore, Maryland, August 25–27, 1993, 1993.
57. Reyna, O.G.P. El Diccionario panhispánico de términos médicos. *Revista de la Sociedad Peruana de Medicina Interna* **2023**, *36*, 169–170.
58. Soto, X.; Perez-de Viñaspre, O.; Labaka, G.; Oronoz, M. Neural machine translation of clinical texts between long distance languages. *Journal of the American Medical Informatics Association* **2019**, *26*, 1478–1487, [<https://academic.oup.com/jamia/article-pdf/26/12/1478/34152142/ocz110.pdf>]. <https://doi.org/10.1093/jamia/ocz110>.
59. Miranda-Escalada, A.; Farré-Maduella, E.; Krallinger, M. Named entity recognition, concept normalization and clinical coding: Overview of the CANTEMIST track for cancer text mining in Spanish, corpus, guidelines, methods and results. In Proceedings of the Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings, 2020.
60. Gonzalez-Agirre, A.; Marimon, M.; Intxaurreondo, A.; Rabal, O.; Villegas, M.; Krallinger, M. Pharmaconer: Pharmacological substances, compounds and proteins named entity recognition track. In Proceedings of the Proceedings of The 5th Workshop on BioNLP Open Shared Tasks, 2019, pp. 1–10.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.